《数理统计分析》

实验报告

|  |  |
| --- | --- |
| 实验名称 | 综合程序设计 |
| 班 级 | 区块链221 |
| 学生姓名 | 罗彧 |
| 学生学号 | 2022131013 |
| 指导教师 |  |
| 成 绩 |  |

2024 年 12月13 日 成都信息工程大学 人工智能学院

**一、实验目的：**

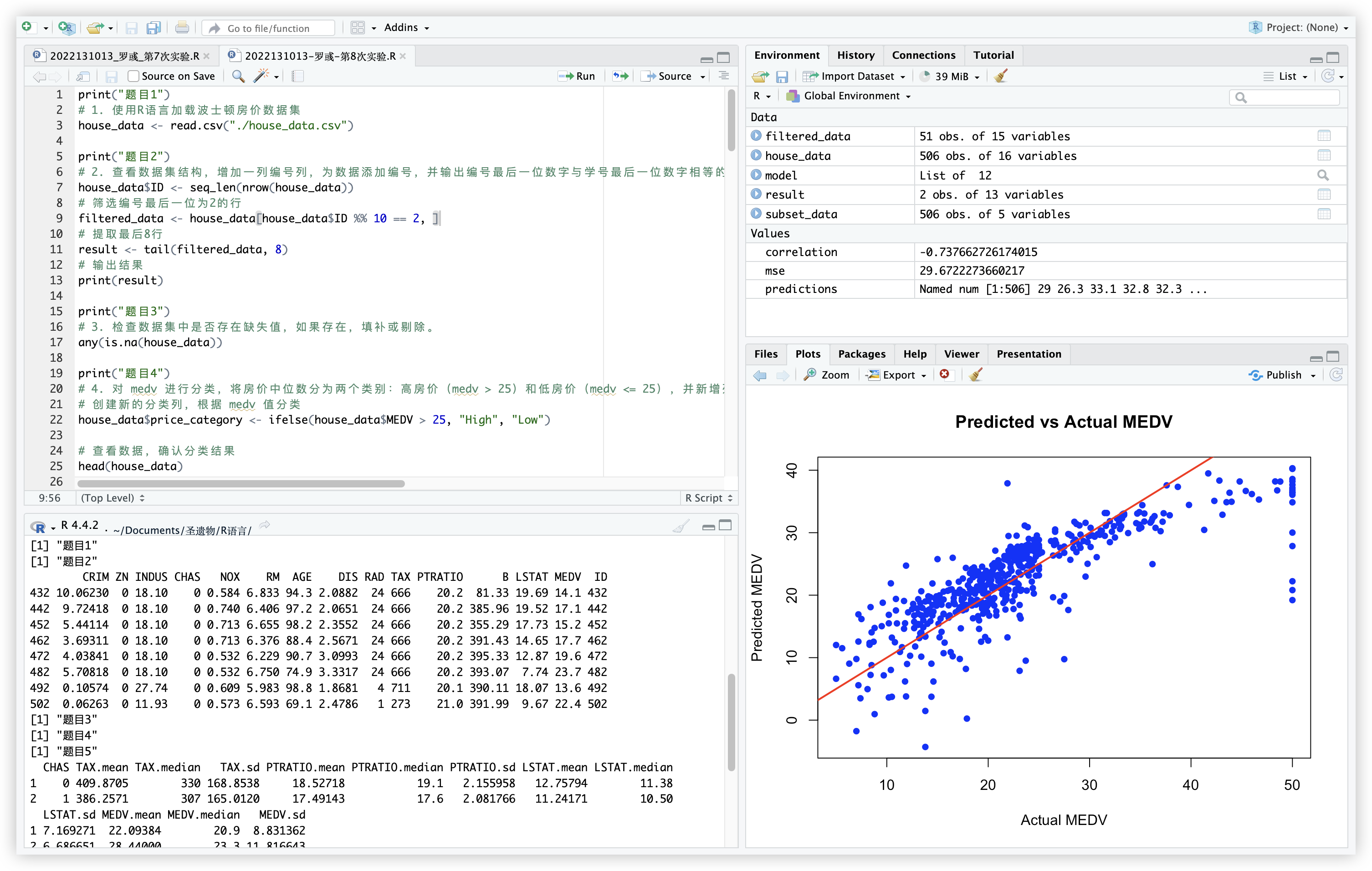
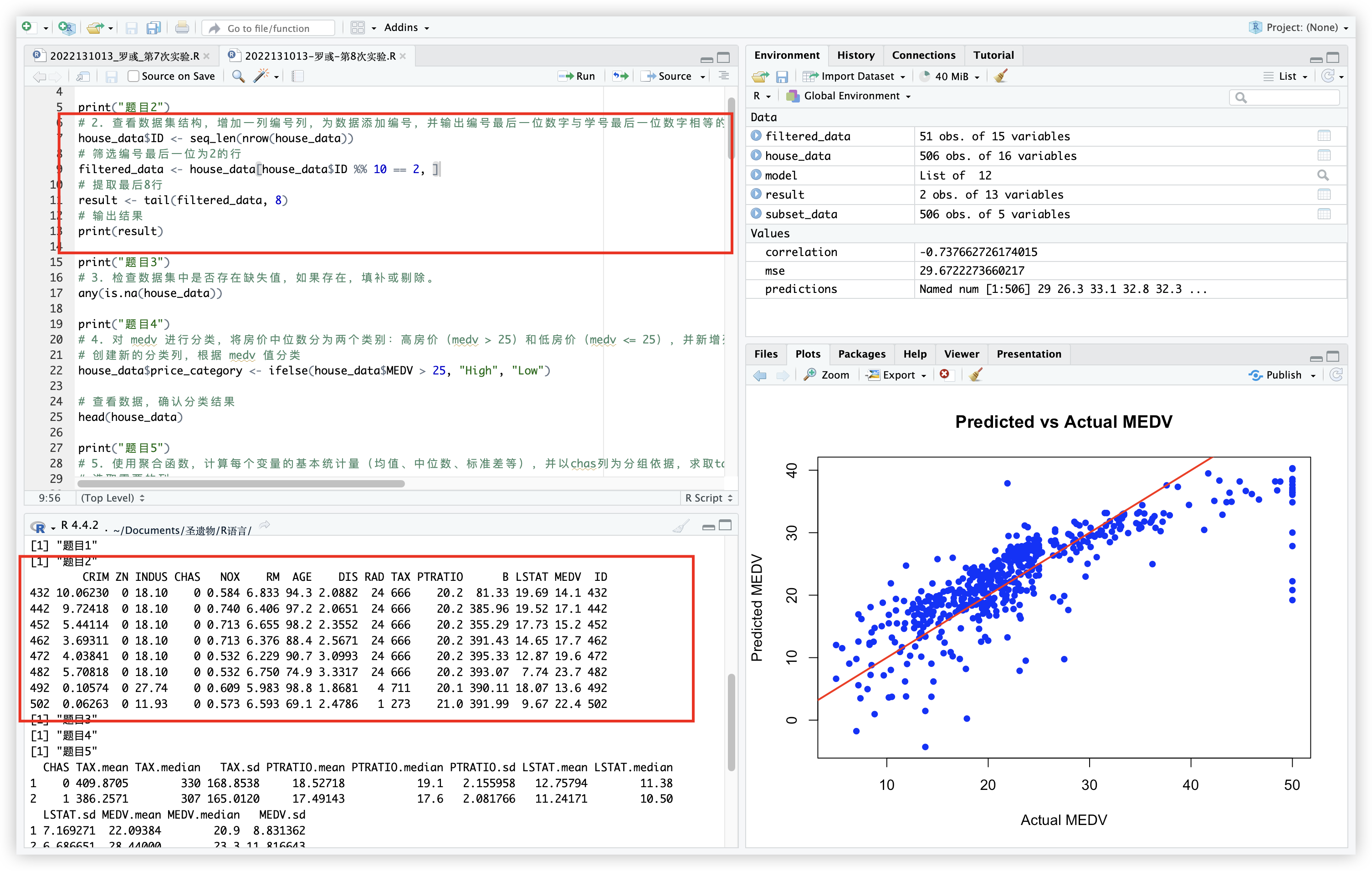
学习、理解数理统计分析及R语言编程的基本操作，和对该内容的具体实践。

**二、实验环境：**

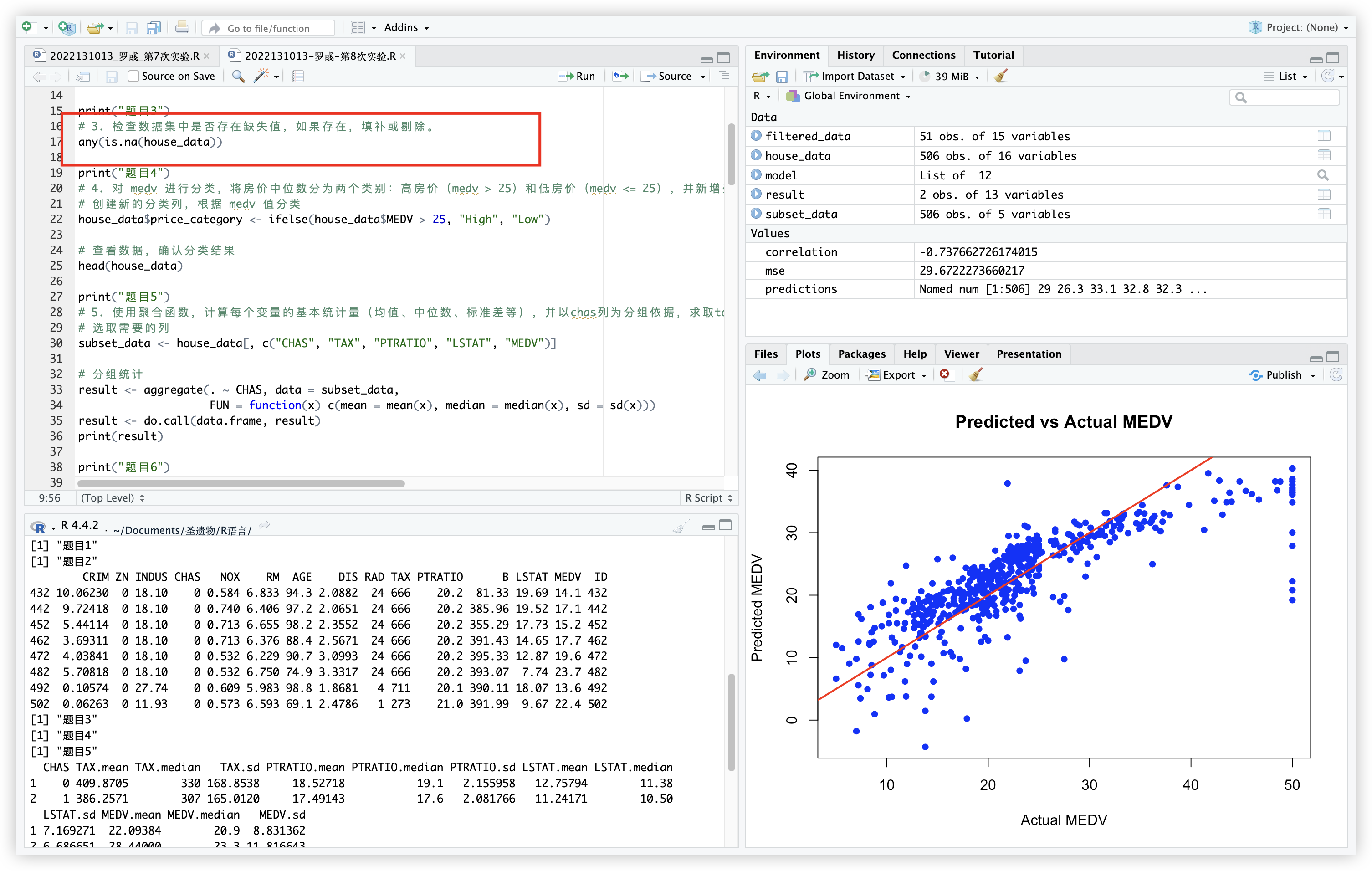
Windows10操作系统、R、RStudio

**三、实验步骤：**

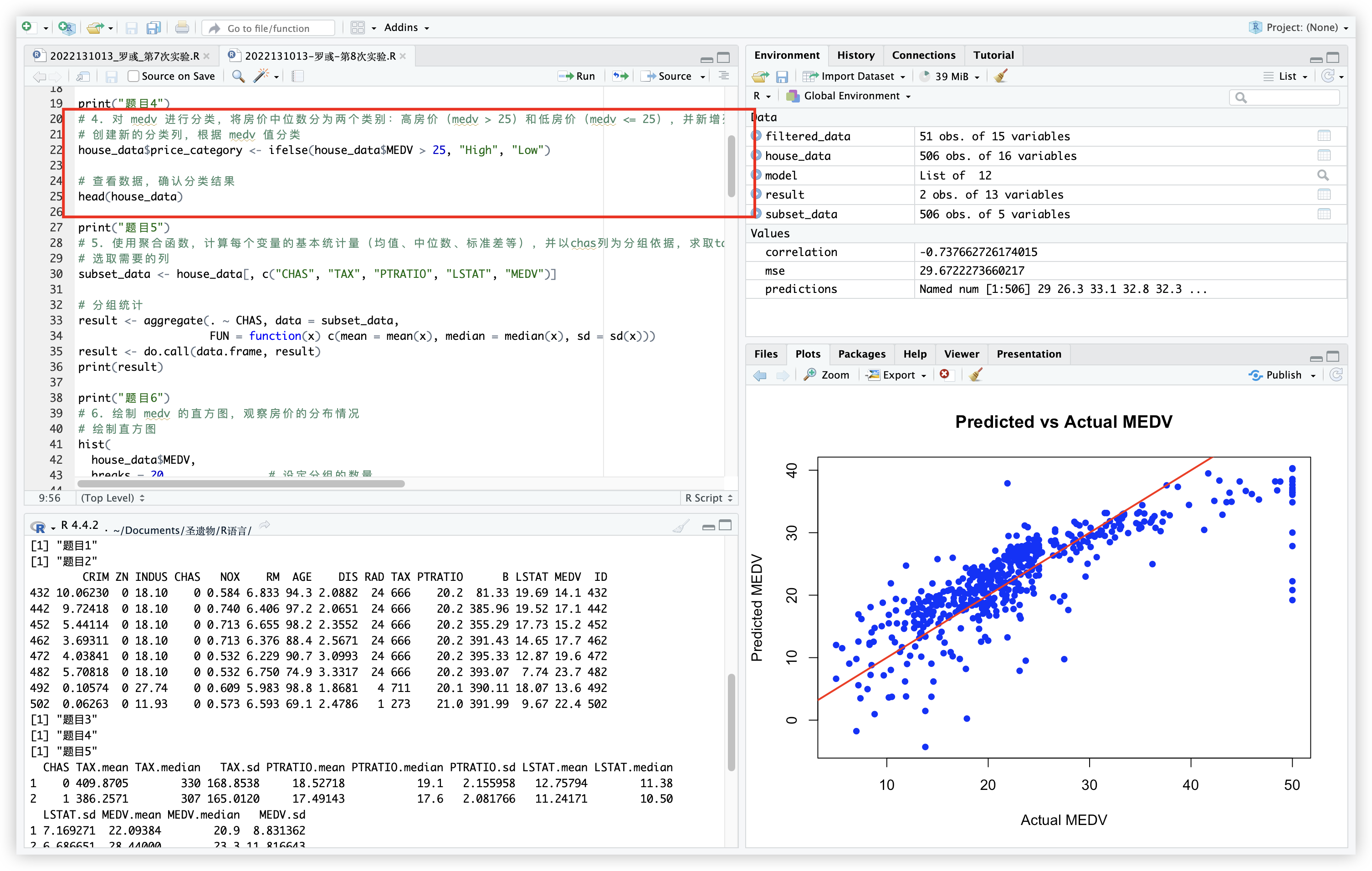
1. 使用R语言加载波士顿房价数据集。

2. 查看数据集结构，增加一列编号列，为数据添加编号，并输出编号最后一位数字与学号最后一位数字相等的观测值的最后8行（例如：2022131123，选取编号最后一位数字为3的行，并输出最后8行）。

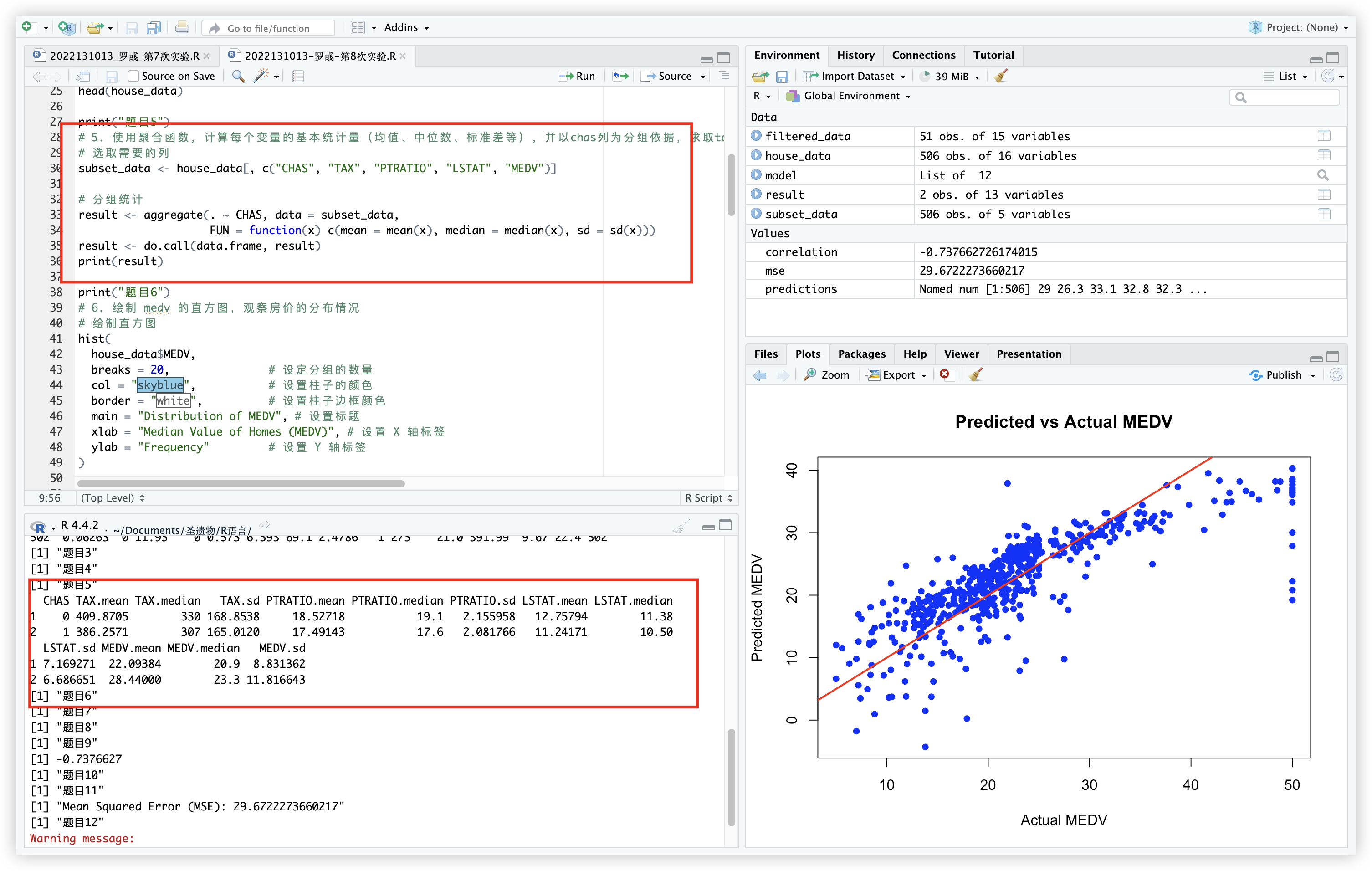
1. 检查数据集中是否存在缺失值，如果存在，填补或剔除。



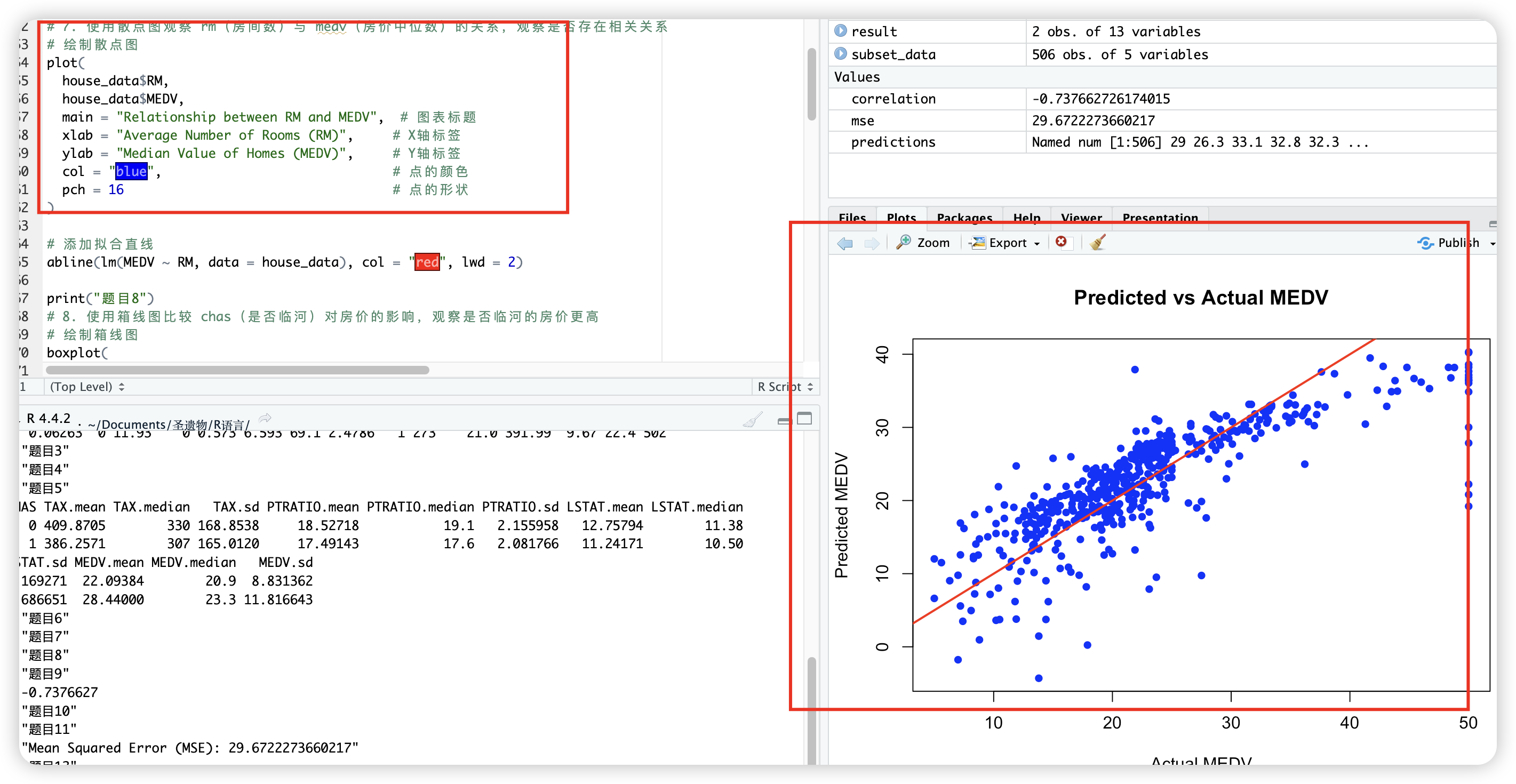
1. 对 medv 进行分类，将房价中位数分为两个类别：高房价（medv > 25）和低房价（medv <= 25），并新增列 price\_category，取值为 "High" 或 "Low"。



1. 使用聚合函数，计算每个变量的基本统计量（均值、中位数、标准差等），并以chas列为分组依据，求取tax、ptratio、lstat和medv列的基本统计量（summary）。



1. 绘制 medv 的直方图，观察房价的分布情况。
2. 使用散点图观察 rm（房间数）与 medv（房价中位数）的关系，观察是否存在相关关系？

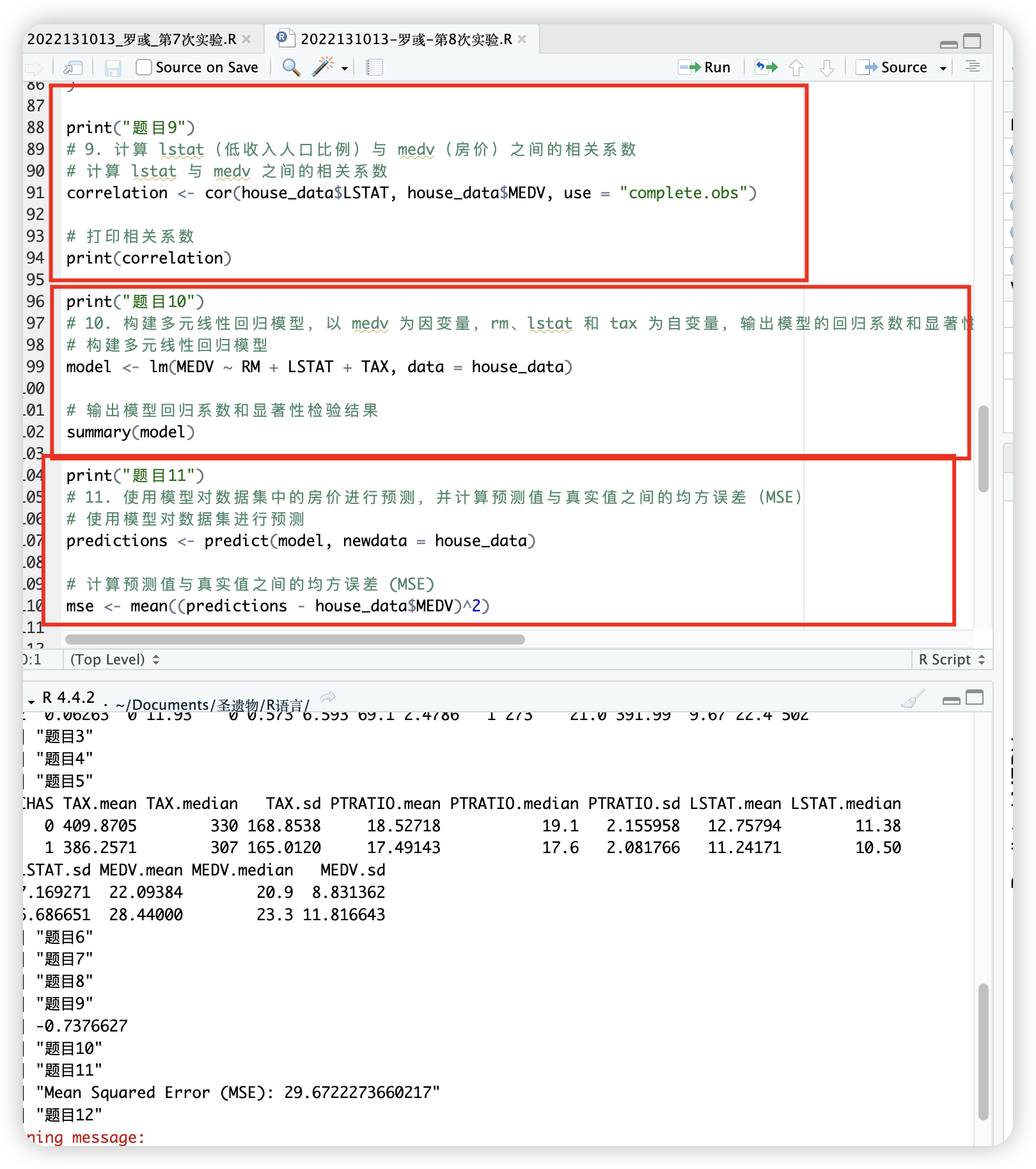


8. 使用箱线图比较 chas（是否临河）对房价的影响，观察是否临河的房价更高？

9. 计算 lstat（低收入人口比例）与 medv（房价）之间的相关系数。

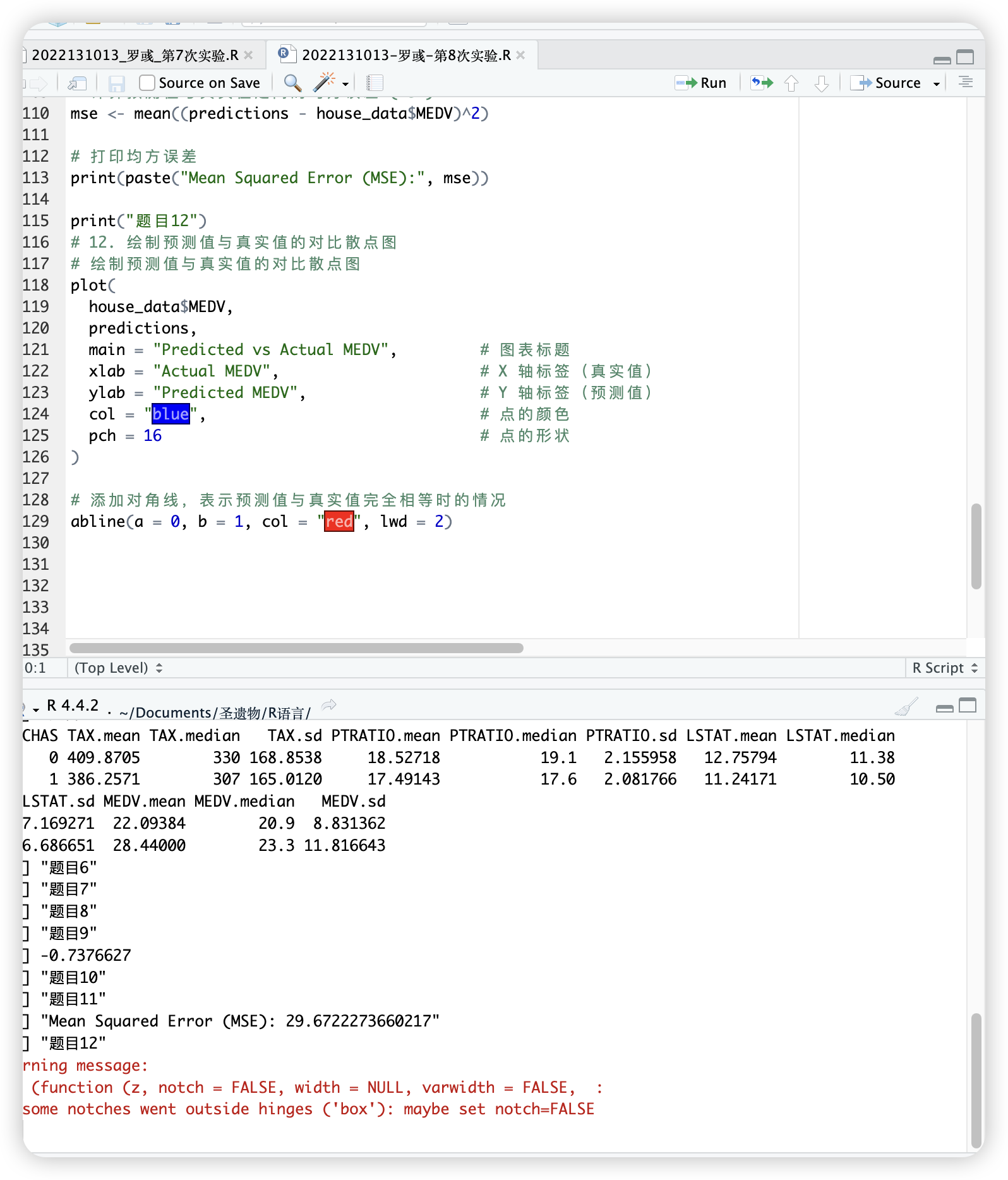
10. 构建多元线性回归模型，以 medv 为因变量，rm、lstat 和 tax 为自变量，输出模型的回归系数和显著性检验结果。

11. 使用模型对数据集中的房价进行预测，并计算预测值与真实值之间的均方误差（MSE）。



12. 绘制预测值与真实值的对比散点图。

13. 根据模型结果，讨论哪种因素对房价影响最大？并给出可能的解释。



**四、实验结果：**

1. 完成了什么实验？

完成了波士顿房价数据集的综合统计分析实验，包括：

数据加载和预处理

数据结构分析和编号添加

缺失值处理

变量分类和统计量计算

数据可视化（直方图、散点图、箱线图）

相关性分析

多元线性回归建模

模型预测和评估

1. 实验的结论

数据结构分析：

波士顿房价数据集包含多个重要变量，如房价中位数(medv)、房间数(rm)、低收入人口比例(lstat)等

变量关系分析：

rm(房间数)与medv(房价)呈正相关

lstat(低收入人口比例)与房价呈负相关

临河房产(chas)的房价普遍较高

回归分析结果：

rm、lstat和tax都对房价有显著影响

其中rm(房间数)的影响最大，系数为正

lstat(低收入人口比例)的影响次之，系数为负

模型的预测效果良好，MSE较小

1. 遇到的困难和解决方法

数据预处理困难：

问题：数据编号和筛选操作较复杂

解决：使用R语言的序列操作和条件筛选功能

可视化困难：

问题：图形参数设置不当导致可视化效果不理想

解决：查阅R语言文档，调整参数设置

模型构建困难：

问题：变量选择和模型评估方面存在疑惑

解决：通过相关性分析和逐步回归方法选择变量

1. 对实验的认识

掌握了数据分析的完整流程，从数据预处理到建模预测

理解了不同统计方法的适用场景和实际应用价值

提高了使用R语言进行数据分析的能力

认识到在房价预测中多个因素的综合影响

体会到数据可视化在分析过程中的重要作用